

W49

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-044506

(43)Date of publication of application : 14.02.1997

(51)Int.Cl.

G06F 17/30

(21)Application number : 07-208554

(71)Applicant : FUJI XEROX CO LTD

(22)Date of filing : 25.07.1995

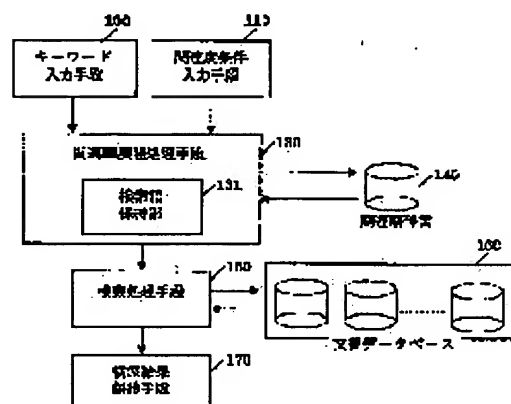
(72)Inventor : KAWAMOTO SHINJI

(54) DOCUMENT RETRIEVAL DEVICE

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain a proper relative word matching the intention of a user and to perform document retrieval operation more efficiently in the document retrieval device which develops a key word into a word related to the key word such as a synonym and retrieves document data by using this relative word, when the document data is retrieved from the key word.

SOLUTION: The key word is inputted through a key word input means 100 and relativity conditions such as the range of relativity of a relative word group to be expanded are inputted through a relativity condition input means 110. A relative word expanding process means 130 extracts the relative word group including the key word from a relative word information storage means 130. Each relative word group as a set of words such as synonyms has a characteristic relativity value showing the degree of relation among the relative words belonging to the group. A relative word expanding process means 130 checks whether or not the relativity of the said extracted relative word group meets the conditions of relativity specified by the relativity condition input means 110. When the conditions are met, a word belonging to the relative word group is used as a retrieval word for retrieval by a retrieval processing means 150.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

w49

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-44506

(43) 公開日 平成9年(1997)2月14日

(51) Int.Cl.⁶

G 0 6 F 17/30

識別記号

庁内整理番号

9289-5L

F I

G 0 6 F 15/403

技術表示箇所

3 2 0 D

審査請求 未請求 請求項の数1 F D (全 14 頁)

(21) 出願番号

特願平7-208554

(22) 出願日

平成7年(1995)7月25日

(71) 出願人 000005496

富士ゼロックス株式会社

東京都港区赤坂二丁目17番22号

(72) 発明者 川本 真司

神奈川県川崎市高津区坂戸3丁目2番1号

K S P R & D ビジネスパークビル 富

士ゼロックス株式会社内

(74) 代理人 弁理士 岩上 昇一 (外2名)

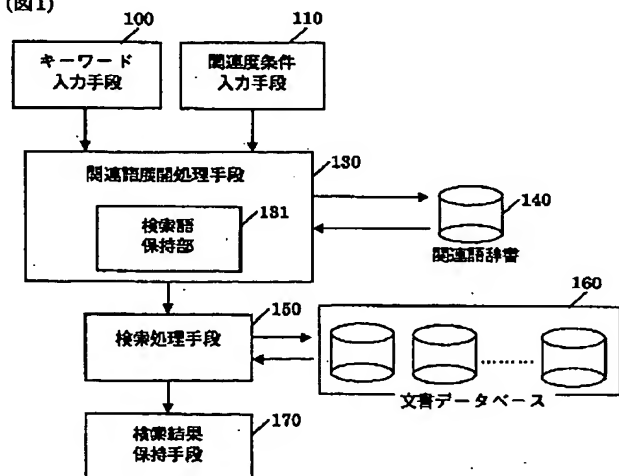
(54) 【発明の名称】 文書検索装置

(57) 【要約】

【課題】 文書データをキーワードにより検索する際、キーワードを同義語、類義語などのキーワードに関連する関連語に展開し、この関連語を用いて検索する文書検索装置において、ユーザの意図に合った適切な関連語を得ることができ、文書検索作業をより効率的に行うこと。

【解決手段】 キーワード入力手段によりキーワードを入力し、関連度条件入力手段により、例えば展開する関連語グループの関連度の範囲などの関連度条件を入力する。関連語展開処理手段は、関連語情報記憶手段からキーワードが含まれる関連語グループを抽出する。同義語、類義語などの語の集まりである関連語グループは関連語グループ単位に、そのグループに属する関連語同士の間連の度合いを示す固有の値の間連度を持っている。関連語展開処理手段は、さらに前記のように抽出した関連語グループの間連度が関連度条件入力手段により指定された間連度の条件を満たすかどうかをチェックする。条件を満たしていればその関連語グループに属する語を検索語として検索処理手段による検索に用いる。

(図1)



【特許請求の範囲】

【請求項 1】 キーワードを入力するキーワード入力手段と、
関連度条件を設定する関連度条件入力手段と、
関連語をグループ化するとともに、各関連語グループに
関連度を関係づけた関連語情報を保持する関連語情報記
憶手段と、
前記キーワード入力手段により入力されたキーワードを
含む関連語グループを前記関連語情報記憶手段から抽出
し、その抽出した関連語グループのうち前記関連度条件
入力手段により入力された関連度条件を満たす関連語グ
ループを求め、その関連語グループに含まれる語を検索
語とする関連語展開処理手段と、
前記関連語展開処理手段により得られた検索語を用いて
文書の検索を行う検索処理手段とを備えた文書検索装
置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明はキーワードを用いて文書
データベース内から所望の文書を検出する文書検索装置
に関する。

【0002】

【従来の技術】 大量の文書が登録された文書データベ
ース内から所望の文書を検索するために、指定したキー
ワードを含む文書を文書データベース内から検出する文書
検索装置が用いられる。キーワードは文書を文書データ
ベースに登録する際に与える方法もあるが、検索をより
柔軟に行うために、あらかじめキーワードを与えず、検
索時にキーワードを自由に指定する全文検索という方法
がある。しかし、ユーザが自由にキーワードを指定でき
るので、検索漏れが発生しやすいという問題があった。
この問題を解消するために、キーワードの類似語、同義
語等キーワードに関連する語も検索語として検索するこ
とにより、検索漏れを減らすという手法がある。キーワ
ードの関連語まで検索すると、検索漏れが少なくなる一
方で、ユーザの検索意図と異なる思わぬ語まで検索語と
して検索されてしまい、余計なものまで検出されてしま
い、所望の文書と検索された文書との適合率が低くな
るという問題があった。このような問題を解決するた
めに、キーワードを関連語に展開する際に展開する関連
語の語数、展開する距離などの条件を設定することによ
り、関連語を制限し、不適切なキーワードによる余計な
検索結果が出ないようにすることが提案されている（例
えば、特開平 5 - 0 2 8 1 9 9 号公報参照）。

【0003】

【発明が解決しようとする課題】 上記従来技術（特開平
5 - 0 2 8 1 9 9 号公報）では、キーワードを関連語に
展開する際に展開する関連語の語数、展開する距離など
を制限する手法を用いているが、この方法では、同じ表
記で複数の意味カテゴリに含まれる語をキーワードとし

て検索した場合、ユーザが意図する語の意味とは異なる
意味カテゴリの語もキーワードの関連語になるため、い
くら展開する関連語の語数などで制限しても、余計なも
のまで検出されてしまうという問題があった。

【0004】 本発明は、文書データをキーワードにより
検索する際、キーワードを同義語、類義語などのキーワ
ードに関連する関連語に展開し、この関連語を用いて検
索する文書検索装置において、ユーザの意図に合った適
切な関連語を得ることができ、文書検索作業をより効率
的に行うことを目的とする。

【0005】

【課題を解決するための手段】 本発明は、キーワードを
入力するキーワード入力手段と、関連度条件を設定する
関連度条件入力手段と、関連語をグループ化するととも
に、各関連語グループに関連度を対応付けた関連語情報
を保持する関連語情報記憶手段と、前記キーワード入力
手段により入力されたキーワードが含まれる関連語グル
ープを前記関連語情報記憶手段から抽出し、その抽出し
た関連語グループのうち前記関連度条件入力手段により
入力された関連度条件を満たす関連語グループを求め、
求めた関連語グループに含まれる語を検索語とする関連
語展開処理手段と、前記関連語展開処理手段により得ら
れた検索語を用いて検索を行う検索処理手段とを備えた
文書検索装置である。

【0006】

【作用】 キーワード入力手段によりキーワードを入力
し、関連度条件入力手段により、例えば展開する関連語
グループの関連度の範囲などの関連度条件を入力する。
関連語展開処理手段は、関連語情報記憶手段からキー
ワードが含まれる関連語グループを抽出する。同義語、類
義語などの語の集まりである関連語グループは関連語グ
ループ単位に、そのグループに属する関連語同士の関連
の度合いを示す固有の値の関連度を持っている。関連語
展開処理手段は、さらに前記のように抽出した関連語グ
ループの関連度が関連度条件入力手段により指定された
関連度の条件を満たすかどうかをチェックする。条件を
満たしていればその関連語グループに属する語を検索語
として検索処理手段による検索に用いる。このように入
力されたキーワードを関連語に展開する際、キーワード
入力とともに関連度条件入力手段により入力された関連
度の条件を満たす関連語グループに属する語のみを関連
語として検出し、検出された語を検索語として文書検索
を行うようにしたので、関係のない文書が検索されるこ
とが従来に比べ少なくなり、検索作業の効率が向上す
る。

【0007】

【実施例】

（第 1 の実施例） 図 1 は、本発明による第 1 の実施例の
文書検索装置の概略の構成を示す図である。この文書検
索装置は、検索のためのキーワードを入力するキーワー

ド入力手段100と、展開する関連語グループの関連度の範囲などの関連度の条件を設定する関連度条件入力手段110と、同義語、類義語など何らかの関連を有する関連語をグループ化するとともに、各関連語グループに関連度を付与した関連語情報を保持する関連語辞書140と、キーワード入力手段100により入力されたキーワードが含まれる関連語グループを関連語辞書140から抽出し、その抽出した関連語グループのうち前記関連度条件入力手段110により入力された関連度条件を満たす関連語グループを求め、求めた関連語グループに含まれる語を検索語とする関連語展開処理手段130と、その関連語展開処理手段130により得られた検索語を格納する検索語保持手段131と、その検索語保持手段131に保持された検索語を用いて文書データベース160に登録されている文書の検索を行う検索処理手段150と、献策処理手段150による検索結果を保持する検索結果保持手段170を備えている。

【0008】図2に示すフローチャートをもとに、以上のように構成された第1の実施例の動作について説明する。まず、キーワード入力手段100からキーワードが入力される(ステップS210)。次に関連度条件入力手段110により、展開する関連語グループの関連度の範囲を入力する(ステップS220)。関連語展開処理手段130により関連語辞書140からキーワードが含まれる関連語グループを抽出する(ステップS231)。同義語、類義語などの語の集まりである関連語グループは関連語グループ単位に固有の値である関連度を持っており、抽出された関連語グループの関連度がステップS220で指定された関連度かどうかをチェックする(ステップS232)。指定された関連度ならば関連語グループに属する語を検索語として検索語保持部131に保持する(ステップS233)。抽出されたすべての関連語グループに対してステップS231～ステップS233の処理を行う。次に検索処理手段150により、検索語保持手段131に保持された検索語の含まれる文書を文書データベース160から検索し、検索結果保持手段170に検索結果を保持する(ステップS240)。

【0009】次に、具体例を用いて実際の処理内容を説明する。例えば『保守』という語をキーワードとして検索する場合を考える。まず、キーワード入力手段100からキーワード『保守』を入力し、関連度条件入力手段110から関連度条件として「関連度2以上」を入力したとする。次に関連語展開処理手段130により、キーワード『保守』を関連語辞書140内の情報をもとに関連語に展開する。関連語辞書の概念図を図3に示す。ある語と関連する語の情報を保持した関連語辞書の構造としてはさまざまな構造が考えられるが、ここに挙げた図3の例では同義語、類義語をまとめた関連語グループ303～307どうしの上位下位関係を定義したシソーラ

ス構造をしている場合を表している。まず、関連辞書内で『保守』という語を含む同義語グループ303、306を検索する。この例では、まず『保守』という語は「原理・主義」というカテゴリ301に属する関連語グループ(ID、22383)303に含まれるので、該関連語グループの情報を抽出する。ここで該関連語グループ303の関連度(DIG)は3で関連度条件として入力された「関連度2以上」を満たすので、該関連語グループ303の語『保守』『与党』『保守党』『保守政党』『勤王』を検索語として検索語保持部131に保持する。また『保守』はカテゴリ「作業・処理」に属する関連語グループ(ID、37573)306にも含まれるが、該関連語グループの関連度(DIG)は1で関連度条件として入力された「関連度2以上」を満たさないため、この関連語グループ306に含まれる語は検索語としては保持されない。このようにしてキーワード『保守』の関連語の中から選出され、検索語保持部131に保持された語を検索語として、検索処理手段150によりデータベースから該検索語を含む文書を検索する。具体的には図4に示す3つの文書を検索対象として考えた場合、従来のキーワードを単に関連語に展開しそれを新しいキーワードとして検索する方式では、3つの文書とも検出されることになるが、本実施例によると、展開されたキーワードの関連語のうちカテゴリ「原理・主義」に含まれる語のみ検索語として検索されるため、文書401、402は検出されるが、文書403は検出されないことになる。

【0010】この第1の実施例によれば、検索もれを少なくするため、キーワード『保守』を関連語展開し、その関連語を検索語として検索する際、指定された関連度条件を満たす関連度グループに含まれる関連語のみを検索語とし、上記の例の場合『修繕』『修理』などの指定された関連度条件を満たさない関連度グループに含まれる関連語は、『保守』の関連語であっても検索語としないことにより、余計な検索結果が検出されることが少なくなり、検索効率が向上する。

【0011】(第2の実施例) この第2の実施例は、第1の実施例において関連度条件の入力方式を変形した例に相当する。図5に第2の実施例の構成を、図6にその動作のフローチャートを示す。図5に示すように関連度条件入力手段510内に関連度グループ表示部511を設けることにより、入力キーワードの属する関連語グループの情報を得た後に、その条件をもとに関連度条件を設定できるようにすることができる。その他の構成は第1の実施例と同じである。

【0012】例えば、『保守』というキーワードを検索する場合、まず、関連語展開処理手段530で入力されたキーワードが属する関連語グループの情報を関連度辞書540から抽出し、その情報を関連語グループ表示部511で表示する(ステップS620)。関連語グルー

ブに関する情報の表示例を図7に示す。この情報をもとにユーザは関連度条件を設定する(ステップS630)。例えば、ユーザが「原理・主義」に関する関連語のみを検索語としたい場合は関連度条件として「関連度が2以上」と設定し、「作業・処理」に関する関連語のみを検索語としたい場合は、「関連度が2以下」と設定する。次に、関連語展開処理手段530でユーザが設定した関連度条件を満たす関連語グループの関連語を検索語として検索語保持部531に保持し(ステップS642)、その検索語をもとに検索処理手段550で検索を行い、検索結果を検索結果保持手段570に保持し(ステップS650)、検索を終了する。

【0013】第2の実施例は、関連度グループ表示部を設け、入力キーワードの属する関連語グループの情報を得た後に関連度条件を設定できるようにしたことにより、関連度条件をより効果的に設定することができ、検索効率の向上につながる。

【0014】(第3の実施例) この第3の実施例は、第1の実施例において関連語グループごとに複数の関連度を持つように構成した例である。図8に本実施例の構成を、図9にその動作のフローチャートを示す。本実施例は、第1の実施例において、検索対象の分野を指定するための検索分野指定手段820を付加するとともに、関連語展開処理手段830に検索分野の情報を検索分野管理テーブルに保持した検索分野管理部832を設けた構成を有する。多種多様の文書が格納された文書内から、所望の文書を検出する際、無駄な検索を少なくするために、例えば、「コンピュータ関連の文書の中からこのような文書を」とか「新聞の記事の中からこのようなものを」という具合に検索分野を指定し、検索範囲を絞りこむ方法がとられることが多い。また、本発明の関連語グループ単位で保持する関連度という値も文書の種類、分野と密接な関係がある。本実施例では、ひとつの関連語グループが分野ごとに複数の関連度を保持する場合の処理について説明する。

【0015】まず、キーワードと関連度条件をそれぞれ、キーワード入力手段800と関連度条件入力手段810によって入力する。ここでは、ユーザが保守政党に関する文書を検索するためキーワードとして『保守』、関連度条件として「関連度2以上」と設定したと仮定する。次に検索分野指定手段820により、検索対象の分野を設定する。この場合、検索分野として「政治／経済関係」を指定する。この検索分野の情報は検索分野管理部832内の検索分野管理テーブルに保持される。図10に検索分野管理テーブルの概念図を示す。分野管理テーブルは、分野番号101と分野名102のフィールドを持ち、分野名が与えられると、対応する分野番号を求めることができる。これらの情報をもとにキーワードを展開処理していく。

【0016】関連語辞書構造の概念図を図11に示す。

この図に示すように各関連語グループ113～117ごとに分野に対応した複数の関連度を保持している。この例では、『保守』という語は「原理・主義」というカテゴリ111に属する関連語グループ(ID、22383)113に含まれるので、該関連語グループ113の情報を抽出する。ここで検索分野として指定された分野の分野番号を検索分野管理テーブル(図10)から検出し、抽出した関連語グループ情報内の複数の関連度の中から該分野番号に対応する関連度をその関連語グループの関連度として展開処理を行う。この例の場合、検索分野として指定された「政治／経済関係」の分野番号は1(図10参照)でそれと対応する関連語グループ情報内の関連度(D-1)は3である。これは関連度条件として入力された「関連度2以上」を満たすので、該関連語グループ113の語『保守』『与党』『保守党』『保守政党』『勤王』を検索語として検索語保持部831に保持する。また、『保守』はカテゴリ「作業・処理」112に属する関連語グループ(ID、37573)116にも含まれるので、該関連語グループ116の情報も抽出する。上記処理と同様な処理を行うと、この関連語グループ(ID、37573)116の関連度(D-1)は1で関連度条件として入力された「関連度2以上」を満たさないで、この関連語グループ116に含まれる語は検索語としては保持されない。つまりキーワード『保守』は「原理・主義」というカテゴリに属する関連語に展開されこの関連語を検索語として、検索処理手段により文書データベースから該検索語を含む文書を検索することになる。

【0017】このように関連度を分野に対応して複数持つことにより、同じキーワードで同じ関連度条件でも検索分野に応じた検索が可能となる。例えば、上記と同様のキーワード『保守』、関連度条件「関連度2以上」で検索した場合でも、検索分野を「コンピュータ関係」で検索すると、この例ではカテゴリ「作業・処理」に属する関連語のみを検索語として検索することになる。また、ここに挙げた例では関連度と検索分野を対応させたが、複数の関連度は検索分野だけでなく、当然その他のいろいろな項目を対応させることが考えられる。例えば、各ユーザIDと関連度を対応させることにより、ユーザごとに別々の展開処理が可能となる。

【0018】このように関連度を分野に対応して複数持つことにより、同じキーワードで同じ関連度条件でも検索分野に応じた検索が可能となる。広い意味では一つの関連語グループに複数の関連度を持てるようにすることにより、ユーザの意図を反映したより細かい関連語展開が行えるため、より正確な検索が行えるようになる。

【0019】(第4の実施例) この第4の実施例は、第3の実施例において関連語辞書の構造を一部変更し、上位概念の関連度を下位概念が受け継ぐように構成した実施例である。意味的が距離の近い関連語グループでは同

じ関連度になる場合が多い。このような関連度の特性を利用して、本実施例では、関連語辞書を関連語グループ間の意味的な上位下位概念を含んだシソーラス構造で作成し、関連語グループ間で共通の関連度は上位のノードで共有するようにする。図12に関連度を共有した場合の本実施例の関連語辞書概念を示す。例えば、関連語グループ(ID. 22383)123と関連語グループ(ID. 22384)124はそのノードに関連度を保持していないので、上位のノード(ID. 22380)121の関連度をその関連語グループの関連度として利用する。関連語グループ(ID. 22385)125は、関連語グループ123や124と意味的には近いが別の関連度であるため、そのノード自身が関連度を保持しており、展開の際にはその関連度を利用する。ID. 22380などの上位のノード121、122においても、意味的に近いノードと関連度を共有できる場合、そのノード自身では関連度を保持せず、さらに上位のノードで関連度を保持することになる。つまり、このような関連語辞書構造の場合、ある関連語グループの関連度はそのノード自身も含めたもっとも近い関連度を保持した上位のノードの関連度ということになる。また、当然のことであるが、図13に示すように、あるノード125'は上位のノード121で関連度の共有部分だけを保持し、各ノードで共有部分との差分だけを持つようにしてもよい。

【0020】第4の実施例によれば、このように意味的な距離の近い関連語グループが同じ関連度を持ちやすい特性を活かし、関連語辞書をシソーラス構造にして関連度を共有することにより、関連度辞書内で同じ情報を重複して持つことがなくなり、メモリを有効利用できる。

【0021】

【発明の効果】従来の方法では意味的に関連性の強い語

も弱い語もすべて同様に検索対象語として検索してしまうため、関係ない多くの文書が候補として検出されてしまい、所望の文書かどうかの判定作業に多くの工数がかかっていたが、本発明によれば、関連性の強さを示す関連度を指定して検索を行うことができるので、この段階で検索の絞りこみが適切に行われ、関係のない文書が候補として検出されることが従来に比べて少なくなり、検索作業の効率が向上する。

【図面な簡単な説明】

【図1】 本発明の第1の実施例の構成図

【図2】 (a) および (b) は、第1の実施例の検索処理の流れを表したフローチャート

【図3】 関連語辞書概念図

【図4】 (a) (b) (c) は、それぞれ検索対象文書の具体例を示す図

【図5】 第2の実施例の構成図

【図6】 (a) および (b) は、第2の実施例の検索処理の流れを表したフローチャート

【図7】 関連語グループに関する情報の表示例を示す図

【図8】 第3の実施例の構成図

【図9】 第3の実施例の検索処理の流れを表したフローチャート

【図10】 検索分野管理テーブルの概念図

【図11】 第3の実施例における関連語辞書の概念図

【図12】 第4の実施例における関連語辞書の概念図

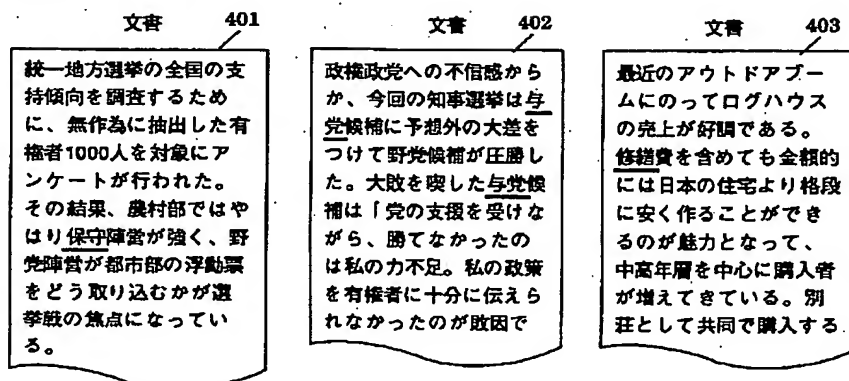
【図13】 図12の関連語辞書の変形例を示す図

【符号の説明】

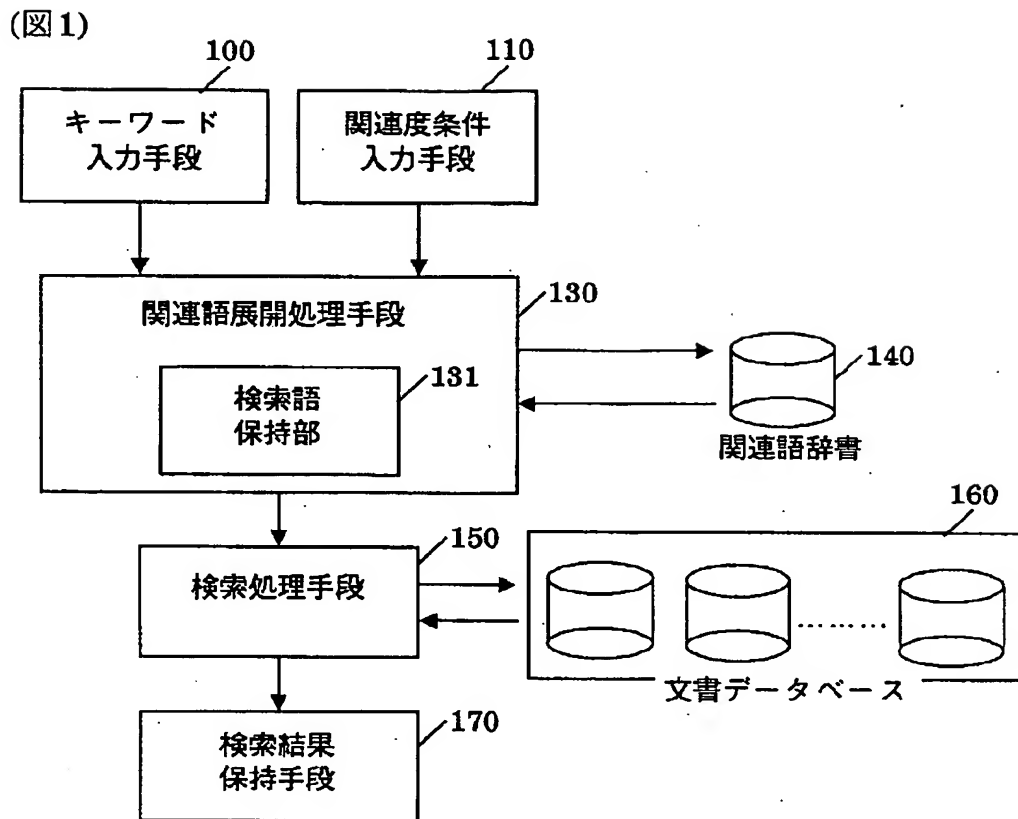
100…キーワード入力手段、110…関連度条件入力手段、130…関連語展開処理手段、131…検索語保持部、140…関連語辞書、150…検索処理手段、160…文書データベース、170…検索結果保持手段。

【図4】

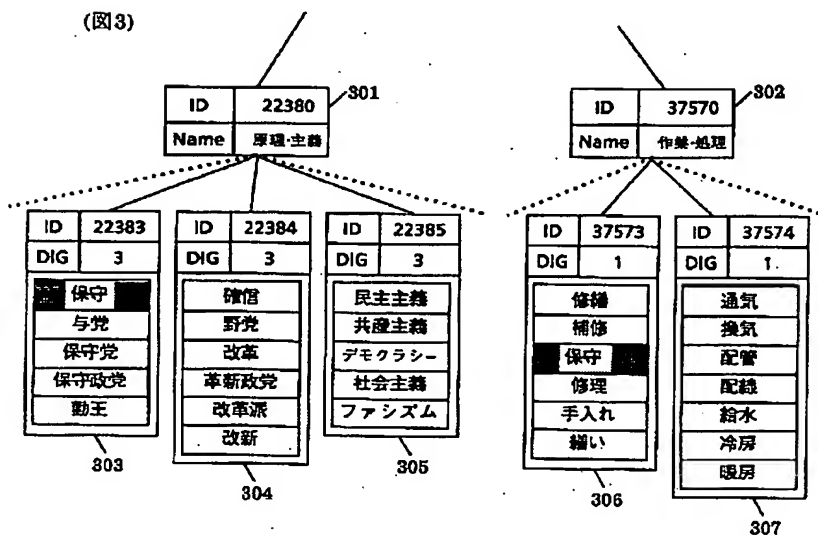
(図4)



【図 1】

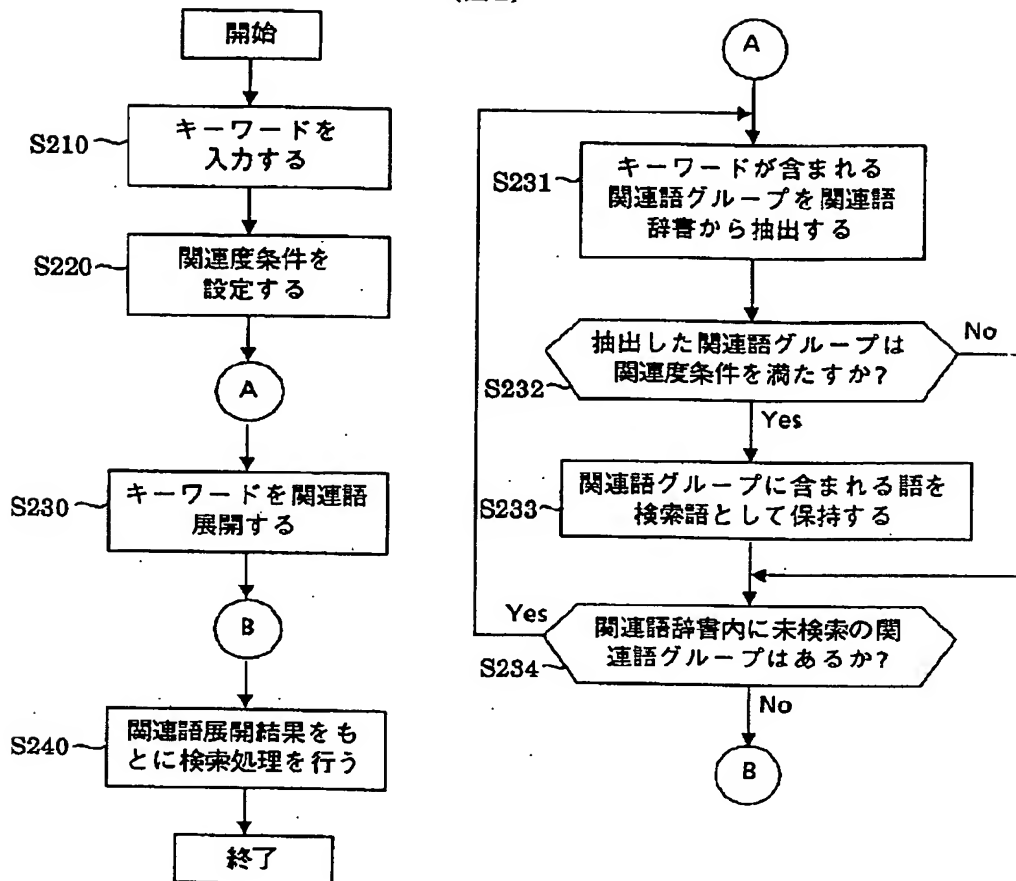


【図 3】



【図2】

(図2)

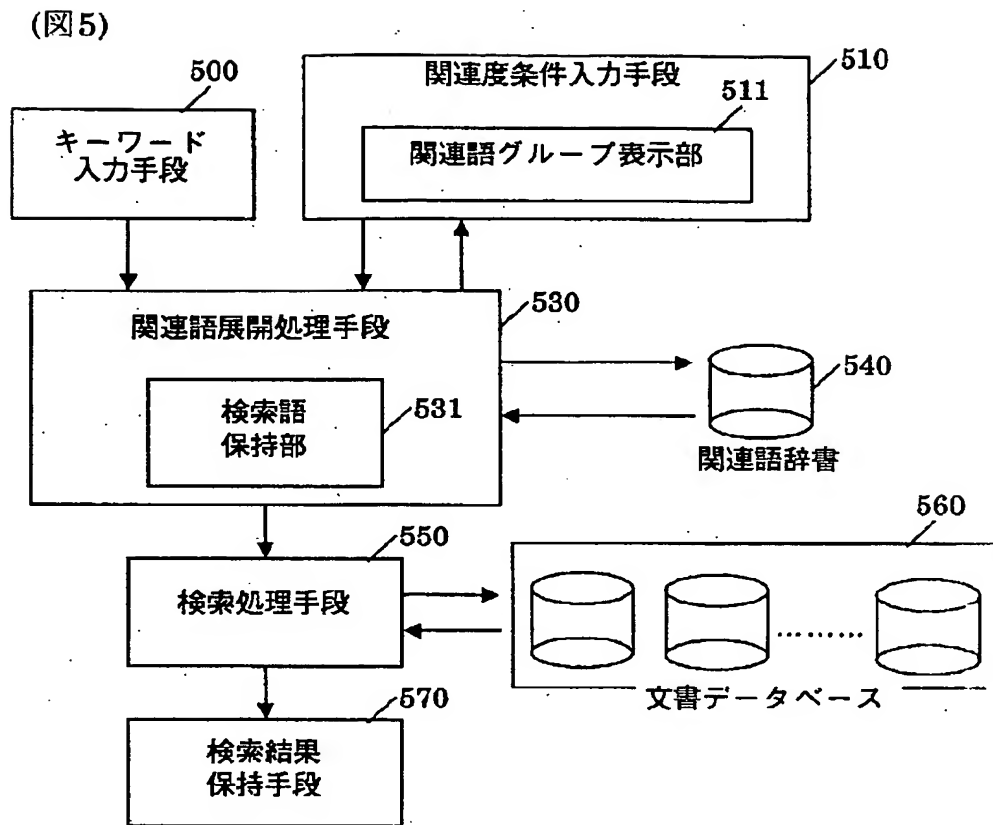


【図7】

(図7)

番号	カテゴリ名	関連度	関連語
1	原理・主義	3	"保守", "与党", "保守党", "保守政党", "勤王"
2	作業・処理	1	"保守", "修繕", "補修", "修理", "手入れ", "繕い"

【図5】



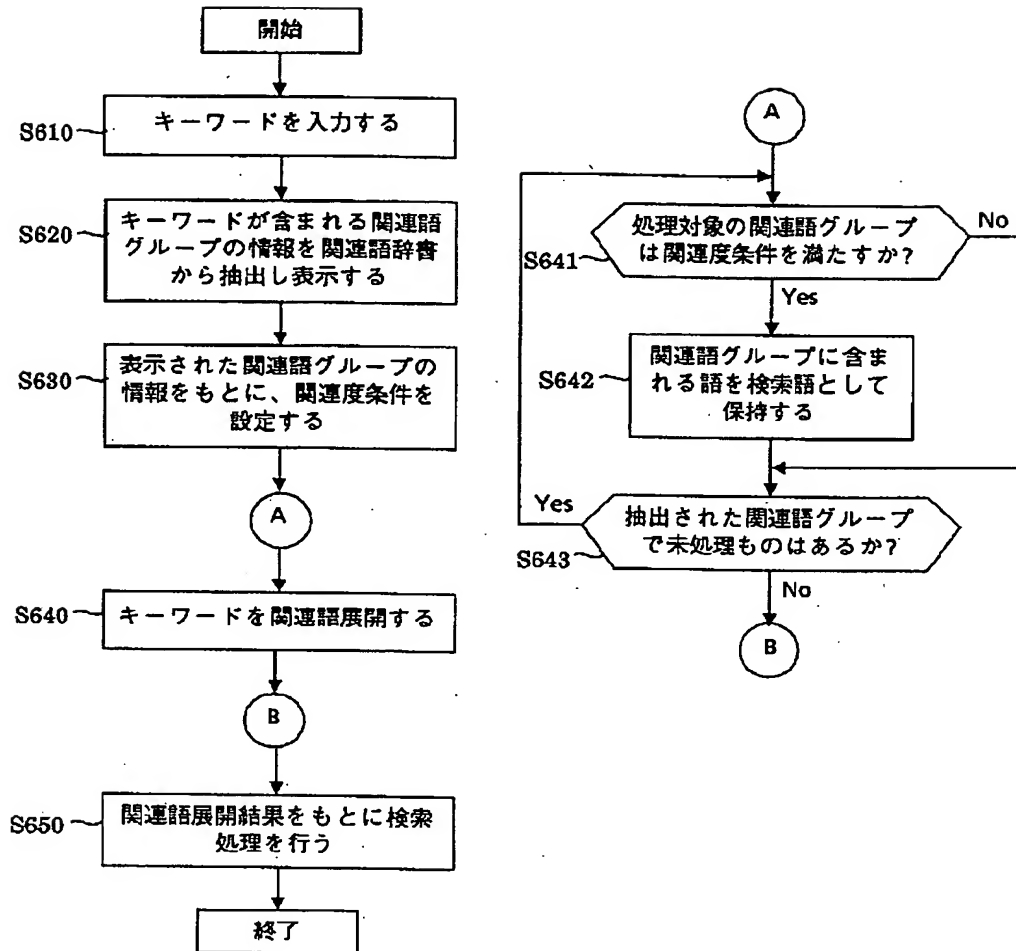
【図10】

(図10)

分野番号	分野名
0	一般文書
1	政治/経済関係
2	医学/科学関係
3	コンピュータ関係
⋮	⋮

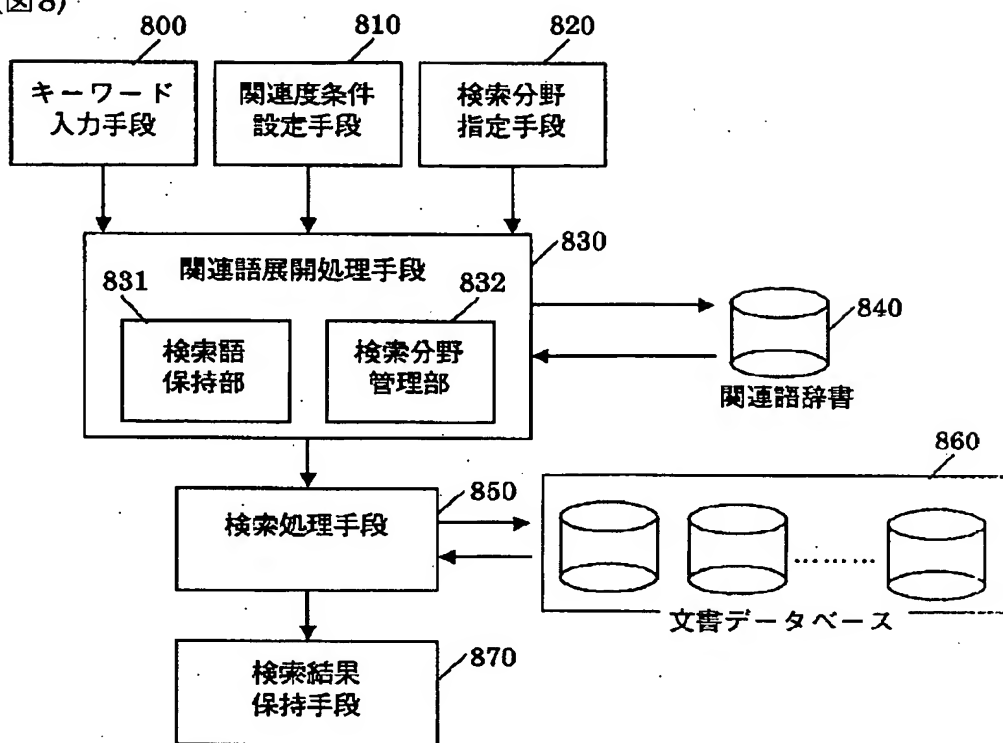
【図 6】

(図 6)

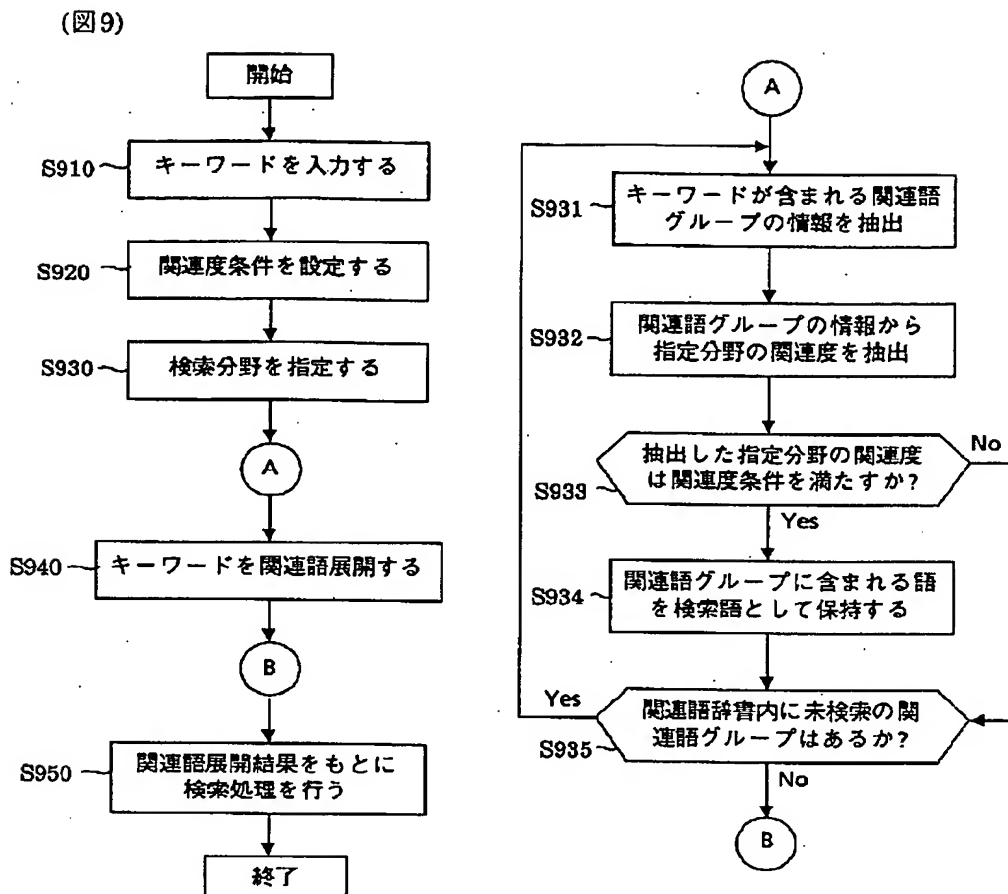


【図 8】

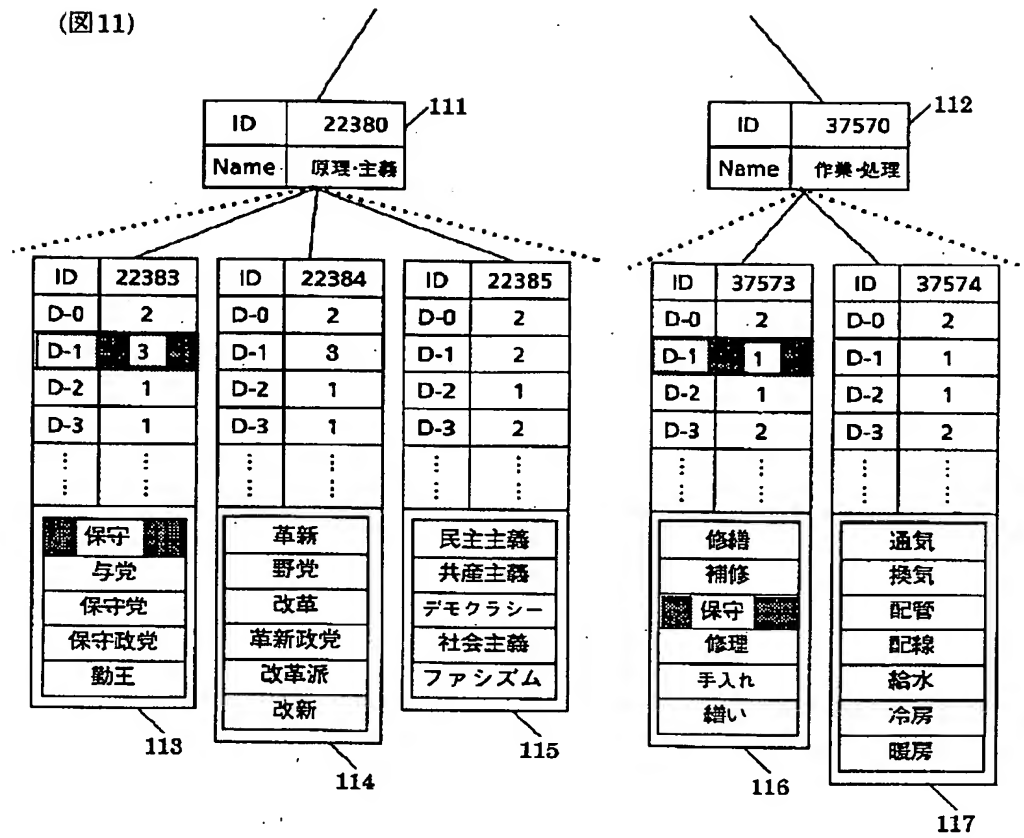
(図 8)



【図9】

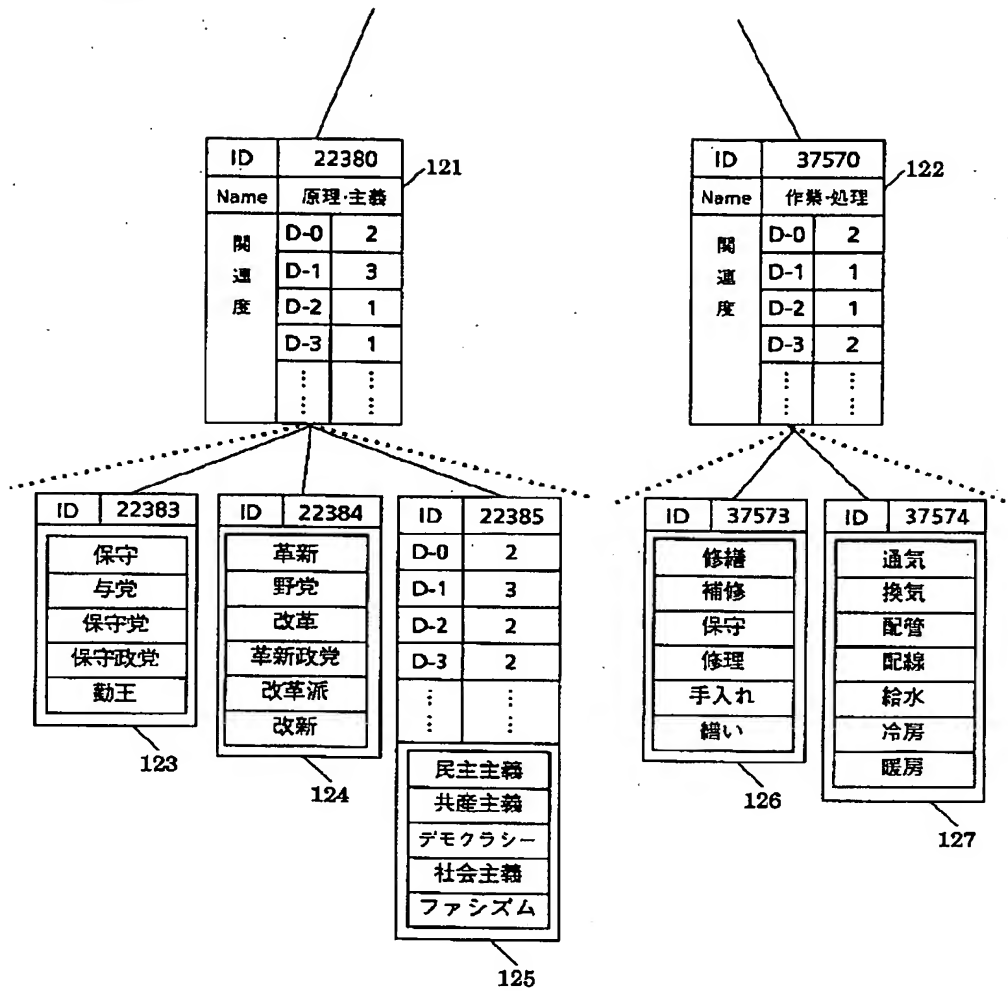


【図 1 1】



【図 12】

(図 12)



【図13】

(図13)

